A Joint Approach for Single-Channel Speaker Identification and Speech Separation

Mowlaee, Pejman; Saeidi, Rahim; Christensen, Mads Græsbøll; Tan, Zheng-Hua; Kinnunen, Tomi; Franti, Pasi; Jensen, Søren Holdt

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASL.2012.2208627

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

A Joint Approach for Single-Channel Speaker Identification and Speech Separation

Pejman Mowlaee, *Member, IEEE*, Rahim Saeidi, *Member, IEEE*, Mads Græsbøll Christensen, *Senior Member, IEEE*, Zheng-Hua Tan, *Senior Member, IEEE*, Tomi Kinnunen, *Member, IEEE*, Pasi Fränti, *Senior Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—In this paper, we present a novel system for joint speaker identification and speech separation. For speaker identification a single-channel speaker identification algorithm is proposed which provides an estimate of signal-to-signal ratio (SSR) as a by-product. For speech separation, we propose a sinusoidal model-based algorithm. The speech separation algorithm consists of a double-talk/single-talk detector followed by a minimum mean square error estimator of sinusoidal parameters for finding optimal codevectors from pre-trained speaker codebooks. In evaluating the proposed system, we start from a situation where we have prior information of codebook indices, speaker identities and SSR-level, and then, by relaxing these assumptions one by one, we demonstrate the efficiency of the proposed fully blind system. In contrast to previous studies that mostly focus on automatic speech recognition (ASR) accuracy, here, we report the objective and subjective results as well. The results show that the proposed system performs as well as the best of the state-of-the-art in terms of perceived quality while its performance in terms of speaker identification and automatic speech recognition results are generally lower. It outperforms the state-of-the-art in terms of intelligibility showing that the ASR results are not conclusive. The proposed method achieves on average, 52.3% ASR accuracy, 41.2 points in MUSHRA and 85.9% in speech intelligibility.**

***Index Terms*—BSS EVAL, single-channel speech separation, sinusoidal modeling, speaker identification, speech recognition.**

Manuscript received December 24, 2011; revised April 12, 2012; accepted June 01, 2012. Date of publication July 13, 2012; date of current version August 24, 2012. The work of P. Mowlaee was supported by the European Commission within the Marie Curie ITN AUDIS under Grant PITNGA-2008-214699. The work of R. Saeidi was supported by the European Community's Seventh Framework Programme (FP7 2007–2013) under Grant 238803. The work of T. Kinnunen was supported by the Academy of Finland (project no 132129). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

P. Mowlaee is with Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum (RUB), 44801 Bochum, Germany (e-mail: pejman.mowlaee@rub.de).

R. Saeidi is with Center for Language and Speech Technology, Radboud University Nijmegen, 6500 HD Nijmegen, The Netherlands (e-mail: rahim.saeidi@let.ru.nl).

M. G. Christensen is with the Department of Architecture, Design and Media Technology, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: mgc@imi.aau.dk).

Z.-H. Tan and S. H. Jensen are with the Department of Electronic Systems, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: zt@es.aau.dk; shj@es.aau.dk).

T. Kinnunen and P. Fränti are with the School of Computing, University of Eastern Finland, FI-70211 Kuopio, Finland (e-mail: tomi.kinnunen@uef.fi; pasi.franti@uef.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2208627

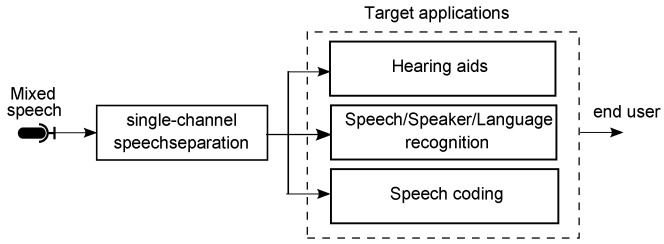

Fig. 1. Block diagram showing how a single-channel speech separation module can be used as a pre-processing stage to enhance the performance of a target application.

I. Introduction

HUMAN beings have the amazing capability of perceiving individual speech sources from mixtures. For machines, however, separating speech mixtures recorded by a single microphone is still a rather difficult task. Designing reliable and robust speech processing systems for adverse conditions is a challenging problem since the observed signal is often corrupted by other interfering signals, making the performance significantly lower compared to that of clean conditions. In extremely noisy environments, a high-quality speech separation algorithm is required as a pre-processing stage prior to the target application, such as hearing aids, automatic speech recognition, speaker/language recognition and speech coding (see Fig. 1). By being able to separate the desired sources from the interfering ones in the mixture, one would expect a better performance in all these applications.

A *single-channel speech separation* (SCSS) system aims at recovering the underlying speaker signals from a mixed signal [1]. At first glance, SCSS is similar to speech enhancement but the goal in SCSS is to recover *all* the underlying signals rather than enhancing the desired speech signal by filtering out the other components. In speech separation, the stronger signal can shift its role to a weaker one at some time-frequency regions, and, further, at different signal-to-signal ratios (SSRs) either one of the signals may dominate the other one. Arguably, one would be interested in separating either of the source signals from their single-channel recorded mixture in certain applications, including signal recovery at low signal-to-noise ratios (SNRs), surveillance and tele-conferencing.

The current SCSS methods can be divided into two major groups, computational auditory scene analysis (CASA) [2], and model-driven methods [3]–[9]. CASA methods use *multi-pitch* estimation methods to extract pitch estimates of the speakers directly from the mixture. The separation performance of CASA-

TABLE I
DIFFERENCES IN MAIN BLOCKS OF EXISTING MODEL-BASED SINGLE-CHANNEL SPEECH SEPARATION. THE PROPOSED ALGORITHMS USED IN THE SYSTEM DIAGRAM IN FIG. 2 ARE HIGHLIGHTED WITH BOLD-FACED FONT

SID and SSR estimation	Spectral feature	Speaker model	Mixture estimator	Signal reconstruction
			Log max [8], [12]–[15]	Ideal binary mask [2]
Iroquois [8], [12], [15]	Gammatone filterbank (GTFB)	Graphical model [8], [15]	MMSE power estimator [16]	Binary mask [13], [14]
Improved Iroquois [5]	Mel-frequency band energy (MFBE)	Factorial HMM [7], [14]	Algonquin [8], [15]	Wiener filter [5], [6], [16]
Closed loop [17]	Log STFT [8], [12], [15]	subband HMM [6]	Maximum likelihood amplitude [9]	Overlap-and-add [7]–[9]
Adapted SID in [18]	**Sinusoidal parameters [9]**	**VQ** [5], **[9]**	**Adapted MMSE in [19]**	**Sq. root Wiener filter in sinusoid [20]**

based methods, as a consequence, is predominantly affected by the accuracy of the multi-pitch estimator, especially when the pitch of one of the speakers is masked by the other [10].

Model-driven methods use pre-trained *speaker models* as *a priori* information to constrain the solution of the ill-conditioned SCSS problem. In particular, source-specific speaker models are incorporated to capture specific characteristics of individual speakers at each frame. As a representative example of model-based methods, non-negative matrix factorization (NMF), decomposes the short-time Fourier transform (STFT) of a mixed signal into a product of two low-rank matrices, namely basis vectors and their corresponding weights [3]. According to [4], NMF cannot always separate speech mixtures when the sources overlap especially when the speakers are of same gender.

The components of a typical model-based SCSS system and algorithms are shown in Table I. SCSS first needs to estimate the identity of underlying speakers and the gain in which the frames are mixed. *Iroquois* [8] is a speaker identification and gain estimation algorithm which uses speaker-specific gain-normalized models to produce a short-list of candidate speakers using the frames dominated by one of the speakers. A modified version of the *Iroquois* system which uses flooring of the exponential argument in likelihood computation obtained slight improvement [5]. Parallel speaker HMMs using Viterbi decoding was used in [11] to identify *only* target speaker which is not enough for model-based speech separation.

The next step is to select a representation of the speech signal which is suitable for separation purpose. Because of the promising results shown in [9], we selected *sinusoidal* features instead of the conventionally used logarithmic short-time Fourier transform (STFT) features [7], [8], [14]. Dynamic models are widely used for speaker modeling [7], [8], [14], [15] due to their great capability to model the sequence of features.

Mixture estimator is a module for finding the best representatives from speaker models to reconstruct mixed-speech frames. It is conventionally performed using log-max model [5]–[8], [14], [15], MMSE power estimator [16] or Algonquin model [8], [15].

The codevectors found by the mixture estimation stage are then passed to *reconstruction* stage which produces the separated signals. In terms of how to reconstruct the separated signals, separation methods are divided into *reconstruction* [7]–[9] and *mask* methods [5], [6], [13], [14], [20]. In the former approach, the codevectors found in the mixture estimation stage are directly used for reconstructing the separated signals. The mask methods, as the name suggests, produce a mask based on the codevectors selected from the speaker models.

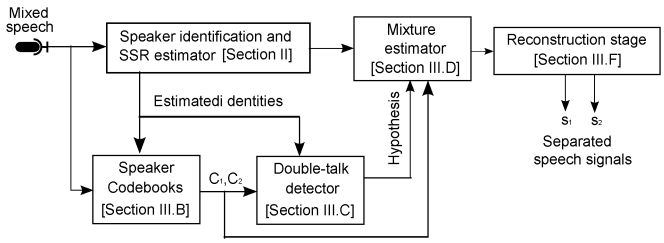

Fig. 2. Block diagram of the proposed joint speaker identification and speech separation system.

The contribution of the current study, as highlighted in Table I and illustrated in Fig. 2, is a novel joint speaker identification and speech separation system. Some of the building blocks were studied individually previously. In addition to the system design, the novel contributions in this paper include extension of the SID module [18] for SSR estimation and generalization of the MMSE mixture estimator in the amplitude domain [19] to sinusoidal features. Considering the high computational complexity of the Iroquois system, a speaker identification (SID) algorithm first proposed in [17] and improved in [18], is utilized in this paper and adopted to the speech separation challenge. Since we look for SCSS algorithm that works equally well also in terms of perceived signal quality basis, the minimum mean square error (MMSE) amplitude spectrum estimation in [19] is adapted for the sinusoidal parametrization. Despite the better upper-bound achieved by dynamic models, we choose static vector quantization (VQ) speaker model which is not limited by the vocabulary and grammar size unlike dynamic models. Moreover, VQ-based models also provide faster decoding. In this work, we use mask-based reconstruction because it leads to promising results in the sinusoidal feature domain [20]. For speaker recognition stage, we use mel-frequency cepstral coefficients (MFCCs) as features and Gaussian mixture models (GMMs) as speaker models and for separation stage we employ sinusoids as features and vector quantization as speaker model.

In evaluating and comparing the proposed method with two state-of-the-art systems [7], [8], we employ a wide range of both subjective and objective quality measures, in addition to standard ASR accuracy. These measures have been introduced in diverse studies in literature but have never been reported together on the speech separation challenge [21]. This has two benefits. Firstly, assessing the separated signals by different metrics rather than ASR has the advantage that the results are expected to carry on to other applications beyond ASR, as indicated in Fig. 1. Secondly, our analysis provides thorough answers to

which of the objective measures correlate best with the subjective measures in SCSS application. The corresponding sections describing each of the presented algorithms are shown inside the blocks in Fig. 2.

A. Speaker Identification and Gain Estimation

Speaker identification (SID) is the task of recognizing speaker identity based on the observed speech signal [22]. Typical speaker identification systems consist of the short-term spectral feature extractor (front-end) and a pattern matching module (back-end). In traditional SID, the basic assumption is that only one target speaker exists in the given signal whereas in *co-channel* SID, the task is to identify two target speakers in a given mixture. Research on co-channel speaker identification has been done for more than one decade [23], yet the problem remains largely unsolved.

Most of the current SCSS systems use the model-driven *Iroquois* system [8] to identify the speakers in a mixed signal. Recognition accuracy as high as 98% on the speech separation corpus [21] has been reported for *Iroquois* [8], which makes it as a viable choice to be used in SCSS systems [7]. In the *Iroquois* system, a short-list of the most likely speakers are produced based on the frames of the mixed signal that are dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* to find the SSR and the two speakers' identities. In subsequent subsections we introduce an alternative approach with lighter computational load in operation phase.

B. Recognition Approach

Generative modeling is widely used for speaker identification [5], [8], [22]. Maximum likelihood (ML) trained GMMs were used in [8]; however, *maximum a posteriori* (MAP) derived GMMs [24] are much more accurate in speaker verification and we follow this latter approach employing conventional MFCCs as feature vectors. Let λ denote a GMM of one speaker. Then the probability density function is

$$p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}). \tag{1}$$

The GMM density function is a weighted linear combination of M Gaussian densities $p_m(\mathbf{x})$, where $p_m(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Here $\boldsymbol{\Sigma}_m$ is a diagonal covariance matrix and the mixture weights w_m further satisfy the constraints $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$. The speaker-dependent GMMs are adapted from a universal background model (UBM) [24]. The UBM is a GMM trained on a pool of feature vectors (MFCCs), extracted from as many speakers as possible, to serve as *a priori* information for the acoustic feature distribution. When adapting the speaker-dependent GMMs, usually only mean vectors are adapted while weights and covariances are shared between all speakers [24].

In traditional speaker recognition, the UBM is trained from a pool of data from different speakers. To characterize mixed speech, in this study we propose to train the UBM (λ_{UBM}) from mixed utterance pairs at different SSR levels. For the ith speaker, the gain-dependent models, λ_{ig}, are adapted from the UBM using ith speaker speech files corrupted by other speakers signal at SSR level g. Using SSR-dependent speaker models, the system captures speaker-specific information when it is contaminated by other speakers. Our method is similar to that of having an SSR-dependent bias in the GMM [8], but we build separate GMMs for each SSR level to utilize the advantages of GMM-UBM system [24]. Using SSR-dependent speaker models enables us to find the most probable speakers along with the most probable SSR level.

1) Frame Level Likelihood Score: One approach to measure the similarity between test utterance and pre-trained speaker models is to calculate frame-level likelihood score. We define the log-likelihood score for a feature vector $\mathbf{x}_t$ given the ith speaker model as $s_{it} = \max_g\{s_{igt}\}$, where

$$s_{igt} = \log p(\mathbf{x}_t|\lambda_{ig}). \tag{2}$$

For each frame we find the most probable speaker. Finding the winner speaker for all of the feature vectors of test utterance, we associate a FLL_{sid} score for each speaker based on the number of frames where the speaker is selected as the winner. During recognition, the UBM is evaluated first and then only the top-scoring Gaussians get evaluated in each SSR-dependent speaker model. We define $\mathrm{FLL}_{\mathrm{ssr}}$ score as the number of times that winner speaker came from g-th SSR-dependent model.

2) Kullback-Leibler Divergence Score: Another approach to measure similarity of the test utterance with speaker models, $\{\lambda_i\}$, is to train a model of the test utterance, λ_e, with MAP adaptation and calculate the distance between λ_e and the speaker models. We use the *Kullback-Leibler divergence* (KLD) as an approximate distance measure between the two probability distributions [25]. Since this distance cannot be evaluated in closed form for GMMs, we use the upper-bound which has successfully been applied to speaker verification [26]:

$$\mathrm{KLD}_{ig} = \frac{1}{2}\sum_{m=1}^{M} w_m(\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig})^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig}). \tag{3}$$

Here g ranges in a discrete set of pre-defined SSR levels, $\boldsymbol{\mu}_{me}$ is the mth mean vector in λ_e and $\boldsymbol{\mu}_{mig}$ is the mth mean vector in λ_{ig}, whereas w_m and $\boldsymbol{\Sigma}_m$ are the weights and the covariances of the UBM, respectively. Considering D as the number of speakers, we form an $D \times G$ distance matrix and associate a KLD_{sid} score for each speaker as the smallest KLD distance (3) over all SSR levels. The original $D \times G$ distance matrix is used as the $\mathrm{KLD}_{\mathrm{ssr}}$ score.

3) Combined Approach: To enable taking benefits from different recognizers, we combine the two scores with equal weights summation. This approach has shown to provide better recognition accuracy than the individual recognizers [18]. Although non-equal weights can be estimated from development data [18], we found that using equal weights yields similar accuracy. Note that we normalize the range of scores from two recognizers before fusion.

C. Selecting the Optimal SID and SSR Pair

The joint speaker identification and separation module produces short-lists of speaker identities and the SSR candidates. In our preliminary speaker identification experiments, we found that the dominant speaker was *always* correctly identified and

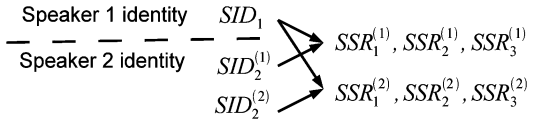

Fig. 3. Demonstration of the reduced search space for speaker-SSR combination. There are $D(D-1)/2 \times G$ possible combination for D speakers and G SSR levels, which is reduced to 2×3 combinations by the proposed joint speaker identification and gain estimation algorithm.

the second speaker also ends up most of the time in the top-3 list. Thus, rather than selecting the top-scoring speaker or the most likely SSR level, we propose the following procedure to select the best pair of speakers and SSR level.

Let SID_1 denote the estimated identity for the first speaker. Assume that the estimated top-2 identities for the second speaker are $\mathrm{SID}_2 = \{\mathrm{SID}_2^{(1)}, \mathrm{SID}_2^{(2)}\}$. Additionally, we define $\mathrm{SSR} = \{\mathrm{SSR}_1^{(i)}, \mathrm{SSR}_2^{(i)}, \mathrm{SSR}_3^{(i)}\}$ as the short-list for SSR candidates consisting of three most likely SSR levels for combination of speakers SID_1 and $\mathrm{SID}_2^{(i)}$ with $i \in \{1,2\}$. The search space is shown graphically in Fig. 3. The speaker identity and SSR candidates in the reduced search space are further passed to the separation module which attempts to reconstruct the mixed signal as combinations of both the two top-scoring speakers and the three SSR candidates. A pair of speakers that minimize the average mixture estimation error (21) in one of the identified SSR-levels (Fig. 3) is selected as the best combination.

II. Single-Channel Speech Separation System

Let $s_z(n)$ denote the nth sample of the observed mixed signal with N samples composed of K additive signals as,

$$s_z(n) = \sum_{k=1}^{K} g_k s_k(n), \quad n = 0, \ldots, N-1. \tag{4}$$

Here, $s_k(n)$ is the kth speaker signal in the mixture, and g_k is its *gain*. Note that the speaker gains, g_1 and g_2, are assumed to be fixed over the entire signal length denoted by N. This assumption, although somewhat unrealistic, is made in most current speech separation systems [21]. For the sake of simplicity and tractability, we consider the case $K = 2$, a mixture of two speakers. We further define $\rho = g_1^2/g_2^2 = 10^{\mathrm{SSR}/10}$ where SSR is the signal-to-signal ratio in decibels. Similar to [27] we assume that the two signals have equal power before gain scaling, i.e., $\sum_{n=0}^{N-1} s_1^2(n) = \sum_{n=0}^{N-1} s_2^2(n) = G_0^2$. By defining $g_z = \sum_{n=0}^{N-1} s_z^2(n)$ and considering $s_1(n)$ and $s_2(n)$ as two independent processes, for large enough N, $E[s_1(n)s_2(n)] = 0$ and $g_z^2 = G_0^2(g_1^2 + g_2^2)$ [27]. The mixed signal can now be represented as below

$$s_z(n) = \frac{g_z\sqrt{\rho}}{G_0\sqrt{1+\rho}} s_1(n) + \frac{g_z}{G_0\sqrt{1+\rho}} s_2(n). \tag{5}$$

The speaker signals $s_1(n)$ and $s_2(n)$ as well as their mixing SSR level (ρ) are unknown while g_z and $s_z(n)$ are given and G_0 is arbitrary for gain scaling.

A. Sinusoidal Signal Representation

The selected features used for separation need to meet at least two requirements: (i) high re-synthesized signal quality, and (ii) low number of features for computational and statistical reasons (curse of dimensionality [28]). A vast majority of the previous separation methods are based on short-time Fourier transform (STFT) features of uniform resolution which poorly match the logarithmic frequency sensitivity of auditory system [12]. In this paper, we choose *sinusoidal* modeling which satisfies both of the aforementioned requirements and leads to improved signal quality compared to the STFT approaches in terms of both objective and subjective measures [9]. Furthermore, in [29], it was shown that applying a sinusoidal coder as speaker model results in a better quantization performance compared to STFT features, in having less outliers [29].

The proposed separation system transforms the underlying speaker signals into a parametric feature set composed of amplitude, frequency and phase vectors of sinusoidal. The sinusoidal parameter estimation is described as follows [9]; On the training data, the STFT magnitude spectrum is calculated using Hann window of 32 msec with hop size of 8 msec. According to the conclusion in [30], replacing the uniform resolution STFT representation with a warped frequency scale, improves the disjointness of the transformed mixtures, and consequently facilitates the separation task since source signals with higher sparsity have less overlap in their mixture. To take the logarithmic sensitivity of the human auditory system into account, we divide the frequency range to frequency bands whose center frequencies are equally distributed on the mel-scale. The frequency bands are non-overlapping and each corresponds to a set of STFT bands. At each band the spectral peak with the largest amplitude is selected. Defining $S_k(\omega)e^{j\phi_k(\omega)} = \mathrm{DFT}_F\{s_k(n)\}$ as the complex spectrum for the kth speaker, with DFT_F as the F-point DFT operator, and $S_k(\omega)$ as its amplitude and $\phi_k(\omega)$ as its phase component, the objective in the sinusoidal parameter estimation used here is to find the set of sinusoids with the following constraints [9]:

$$\omega_{k,i} = \arg\max_{\omega \in \Omega_i} S_k(\omega), \tag{6}$$

$$A_{k,i}e^{j\phi_{k,i}} = S_k(\omega_{k,i})e^{j\phi_k(\omega_{k,i})}, \tag{7}$$

where Ω_i is a set composed of all discrete frequencies within the ith band and $i \in [1, L]$ with L the number of frequency bands (sinusoidal model order), and $A_{k,i}, \omega_{k,i}, \phi_{k,i}$ as the amplitude, frequency and phase for the ith sinusoid, respectively, and $argmax(\cdot)$ returns the argument where $S_k(\omega)$ attains its maximum value. It should be noted that as L approaches to F, each frequency subband include one DFT point.

Assume that the kth speaker time-domain signal is denoted by $\{s_k(n)\}_{n=0}^{N-1}$ where $k \in [1, 2]$, n as the time sample index and N as the window length in samples. For $n = 0, \cdots, N-1$, at each frame, we represent $s_k(n)$ as [31]

$$s_k(n) = \sum_{i=1}^{L} A_{k,i} \cos(n\omega_{k,i} + \phi_{k,i}) + e_k(n), \tag{8}$$

where $e_k(n)$ is the estimation error, i is an index that refers to the ith sinusoidal component. The sinusoidal components are characterized by the triple set $[\boldsymbol{\alpha}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k]$ denoting the amplitude, frequency and phase. We define $\boldsymbol{\alpha}_k = [A_{k,1} A_{k,2} \cdots A_{k,L}]^T$, $\boldsymbol{\omega}_k = [\omega_{k,1} \omega_{k,2} \cdots \omega_{k,L}]^T$, $\boldsymbol{\phi}_k = [\phi_{k,1} \phi_{k,2} \cdots \phi_{k,L}]^T$ as the kth speaker's amplitude, frequency and phase vectors, respectively, each of size $L \times 1$, and L being the sinusoidal model order. We further define

$$C(\boldsymbol{\alpha}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k) = \left| \mathrm{DFT}_F \left\{ \left(\sum_{i=1}^{L} A_{k,i} \cos(n\omega_{k,i} + \phi_{k,i}) \right) w(n) \right\} \right| \tag{9}$$

where $C(\boldsymbol{\alpha}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k)$ is the amplitude spectrum of the kth source represented by the triple $[\boldsymbol{\alpha}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k]$ of size $3L \times 1$, and $w(n)$ is a window function. For a single speaker, the interference effects by sinusoids, taken per frequency subbands, are negligible as the frequencies are rather well separated with respect to each other. Then, from Fourier transformation, the power spectrum for the harmonic-part for the kth source is well approximated by $P_k(\omega) \approx \sum_{i=1}^{L} A_{k,i}^2 W(\omega - \omega_{k,i})$ where $W(\omega)$ is the power response of the Fourier transform for window function, $w(n)$. The magnitude of the STFT is then approximated by $S_k(\omega) \approx \sum_{i=1}^{L} A_{k,i} W(\omega - \omega_{k,i})$ as in [32].

Taking the highest peak of the amplitude spectrum in (6) is equivalent to choosing the maximum likelihood estimate for frequency of single sinusoid in white Gaussian noise per band [33, ch. 13]. In case of no peak detection in a frequency band, we assign an insignificant value of 0.001 for the amplitude and assign the band's center frequency as the frequency of the sinusoid. According to our previous studies [9], [29], this choice would not change the perceived quality of the reconstructed speech but helps to avoid the complicated variable dimension VQ by preserving the fixed dimensionality of the sinusoids.

B. Speaker Codebooks

Split-VQ codebooks composed of sinusoidal amplitude and frequency vectors are used as speaker models [9], [29]. In the split-VQ codebooks, each amplitude vector have several corresponding frequency vectors. The training stage to obtain split-VQ codebooks is composed of two steps; First the amplitudes of sinusoids are coded, then as the second stage, frequency codevectors are found by using vector quantization on frequency candidates assigned to each amplitude codeword found in the first step. For more details see [29]. At the end of the training stage, the codebook entries composed of amplitude and frequency parts are both of the same dimensionality as the sinusoidal model order (L). The split-VQ used in this paper can be replaced by any other sinusoidal coder already available in the speech coding literature, e.g., [34]. The importance of the quantization step is explained in Section IV-C-3.

C. Double-Talk Detection

A mixed speech signal can be classified into single-talk (one speaker), double-talk (speech mixture), and noise-only regions. This information can be used to simplify the computationally expensive separation task since we only need to process the mixed frames with the separation system. To detect double-talk regions with two speakers present, we employ a MAP detector proposed recently in [35]. The proposed method is based on multiple hypothesis test and can be implemented in both speaker-dependent and speaker-independent scenarios. We consider here the speaker-dependent scenario since the information for speaker identities are given by SID module (Section II). We use three candidate models for describing the mixed signal, namely,

$\mathbf{M}_0$: None of the speakers are active (non-speech)
$\mathbf{M}_1$: One of the speakers is active (single-talk)
$\mathbf{M}_2$: Both of the speakers are active (double-talk)

We use the decision making among $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ to narrow down the separation problem only for the mixed frames. For the single-speaker frames, the observed signal is directly re-synthesized according to the corresponding speaker models. For more details of the method, refer to [35].

D. Sinusoidal MMSE Estimator for Mixture Amplitude

In model-driven speech separation we estimate the codevectors in the speaker models whose combination best matches the mixed signal. This is accomplished by employing a *mixture estimator*. In the following, we present the MMSE mixture estimator for the SCSS problem. We define $S_z(\omega) e^{j\phi_z(\omega)} = \mathrm{DFT}_F\{s_z(n)\}$ as the complex spectrum for the mixture. Beginning from the relationship between the mixed signal and the underlying signals in time-domain given in (4), we have

$$S_z(\omega) = \sqrt{g_1^2 S_1^2(\omega) + g_2^2 S_2^2(\omega) + 2 g_1 g_2 S_1(\omega) S_2(\omega) \cos\theta(\omega)}, \tag{10}$$

where we define $S_1(\omega)$, $S_2(\omega)$ and $S_z(\omega)$ are the frequency components of the magnitude spectrum for the first speaker, the second speaker and the mixed signal, respectively. We also define $\theta(\omega) = \phi_1(\omega) - \phi_2(\omega)$ as the phase difference between the kth frequency bin of the underlying spectra. Dividing both sides of (10) by $g_1^2 S_1^2(\omega) \neq 0$, we arrive at

$$\frac{S_z^2(\omega)}{g_1^2 S_1^2(\omega)} = 1 + \frac{g_2^2 S_2^2(\omega)}{g_1^2 S_1^2(\omega)} + \frac{2 g_1 g_2 S_1(\omega) S_2(\omega)}{g_1^2 S_1^2(\omega)} \cos\theta(\omega). \tag{11}$$

By defining $\tilde{S}_z(\omega) \triangleq \ln S_z^2(\omega)$ and $\tilde{S}_i(\omega) \triangleq \ln S_i^2(\omega)$ for $i = \{1, 2\}$ and using (11) we get

$$\tilde{S}_z(\omega) = \ln \frac{g_z^2 \rho}{G_0^2 (1+\rho)} + \tilde{S}_1(\omega) + \ln\left(1 + \frac{1}{\rho} e^{\tilde{S}_2(\omega) - \tilde{S}_1(\omega)}\right) + \ln\left(1 + \frac{\cos\theta(\omega)}{\cosh\left(\frac{-\ln\rho + \tilde{S}_2(\omega) - \tilde{S}_1(\omega)}{2}\right)}\right). \tag{12}$$

A similar expression can be derived by dividing both sides of (10) by $S_2^2(\omega) \neq 0$. The derivation presented here is similar to [36], for representing the relationship among the log-spectra of the noisy signal for speech enhancement, but adopted here for speech mixture of two speakers.

In the following, we derive a closed-form representation for the MMSE mixture estimation in sinusoids. Integrating out the mixture phase modeled with uniform distribution [37], the mixture magnitude spectrum domain is given by

$$\hat{S}_z(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{0.5 \tilde{S}_z(\omega)} d\theta(\omega). \tag{13}$$

where $\hat{S}_z(\omega)$ is the sinusoidal MMSE estimate for mixture magnitude spectrum averaging out $\theta(\omega)$ when we replace the kth speaker signal spectrum with its estimated spectrum represented by its sinusoidal features denoted by $C(\boldsymbol{\alpha}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_z)$ where $\boldsymbol{\alpha}_k$ and $\boldsymbol{\omega}_k$ are calculated using (5) and (6). It is important to note that, as we have no access to each speaker's phase value $\{\boldsymbol{\phi}_k\}_{k=1}^2$ in its corresponding sinusoidal representation, we set $\boldsymbol{\phi}_k = \boldsymbol{\phi}_z$. This choice is in line with the fact that the phase of the noisy observation is the MMSE phase estimate for the clean speech [38]. Furthermore, the authors in [32] showed that the choice of the phase spectrum sampled at frequencies of sinusoids as the estimated phase of sinusoids is sufficient for estimating the sinusoidal parameters in MMSE sense. Following a similar approach as in [19], (13) simplifies to (14) shown at the top of the page,

$$\begin{aligned}\hat{S}_z(\omega) &= f\left(C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_z), C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_z), \rho\right) \\ &= \left[\frac{\sqrt{\rho}}{\sqrt{1+\rho}} C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_z) + \frac{1}{\sqrt{1+\rho}} C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_z)\right] \\ &\quad \times \frac{g_z \mathcal{E}\left(\gamma_\rho(\omega)\right)}{\pi G_0}, \end{aligned} \tag{14}$$

where $f(\cdot)$ is the MMSE mixture approximation and $\gamma_\rho(\omega) = 2/(\sqrt{\xi_\rho(\omega)} + (1/\sqrt{\xi_\rho(\omega)}))$ and we define $\mathrm{SSR}_{\mathrm{prior}} \triangleq \xi_\rho(\omega) = \sqrt{\rho} C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_z)/C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_z)$ and $\mathcal{E}(\cdot)$ is the complete Elliptic integral of the second kind. This integral can be approximated by the following series:

$$\mathcal{E}(\eta) = \pi \left\{1 - \sum_{m=1}^{\infty} \left[\prod_{v=1}^{m} \left(\frac{2v-1}{2v}\right)^2\right] \frac{\eta^{2m}}{(2m-1)}\right\}. \tag{15}$$

The Elliptic series denoted by $\mathcal{E}(\cdot)$ can also be written as

$$\mathcal{E}\left(\gamma(\omega)\right) = \frac{\pi}{2} {}_2F_1\left(-0.5, 0.5; 1; \gamma^2(\omega)\right) \tag{16}$$

where ${}_2F_1(a, b; c; t)$ is Gauss' hypergeometric function with t as an argument replaced by $\gamma^2(\omega)$. Provided that $|t| \leq 1$, $\mathcal{E}(\gamma(\omega))$ will converge absolutely, and since $\gamma(\omega) \leq 1$, convergence is indeed guaranteed. Note that the values of ${}_2F_1(\cdot)$ can be found from a look-up table since it depends on a single variable, $\gamma(\omega)$. This helps keep the complexity of the mixture estimator low.

Previous separation systems used either *max-model* [14] or *Algonquin model* [8] as their mixture estimator. A simplified version of the max-model, MAX-vector quantization (MAX-VQ) was used in [5], [7], [13]. In [8], both the Algonquin and the max-model were studied and compared, and Algonquin was found to perform slightly better. The max model and *Algonquin model* use MMSE criterion in log-power and power spectrum domain considering the phase as a random variable. The proposed mixture estimator also takes this into account in amplitude domain. Furthermore, according to [15], specifying the mixture estimation stage in the log spectral domain is convenient because speech states can be represented efficiently as a mixture of Gaussians in the log-spectrum. For reconstruction purposes, then, they use anti-logarithmic transformation. In this paper, we solve the problem directly in the spectrum amplitude domain matched with our signal reconstruction stage (see Section III-F), without the logarithmic mapping.

E. Estimating Optimal Codebook Indices

Here, we explain how to find the estimated mixture magnitude spectrum, $\hat{S}_z(\omega)$ given in (14), at each frequency bin. To implement the mixture estimator in (14), we need the spectra of the two speakers, $C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_1)$ and $C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_2)$. In the expression for MMSE estimate for mixture amplitude in (14), the signal spectra of the underlying speakers were considered to be given. However, in the experiments, we relax this assumption by choosing their estimates as $C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_1)$ and $C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_2)$ selected from the pre-trained codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$ of the two speakers. The estimates for $C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1)$ and $C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2)$ are obtained from the codebooks of the two speakers, $\mathbb{C}_1 = \{\mathbf{c}_1^{(1)}, \mathbf{c}_2^{(1)}, \ldots, \mathbf{c}_r^{(1)}, \ldots, \mathbf{c}_M^{(1)}\}$ and $\mathbb{C}_2 = \{\mathbf{c}_1^{(2)}, \mathbf{c}_2^{(2)}, \ldots, \mathbf{c}_q^{(2)}, \ldots, \mathbf{c}_M^{(2)}\}$, respectively, where $\mathbf{c}_r^{(1)}$ and $\mathbf{c}_q^{(2)}$ refer to the rth and qth codevector in the codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively. Let $S_z(\omega)e^{j\phi_z(\omega)} = \mathrm{DFT}_F\{s_z(n)\}$ to be the discrete Fourier transform of the mixture. Each codebook consists of a pair of amplitude and frequency ($\{\boldsymbol{\alpha}, \boldsymbol{\omega}\}$), and M is the number of codevectors in the speaker models [29].

Let $\mathbf{e}^c(\omega)$ be the full-band mixture estimation error in complex spectrum domain defined as the error difference between the complex spectrum of mixture, $S_z(\omega)e^{j\phi_z(\omega)}$ and the estimated complex spectrum of the mixture, $\hat{S}_z(\omega)e^{j\hat{\phi}_z(\omega)}$ as follows:

$$\mathbf{e}^c(\omega) = S_z(\omega)e^{j\phi_z(\omega)} - \hat{S}_z(\omega)e^{j\hat{\phi}_z(\omega)}. \tag{17}$$

At each frequency subband $i \in [1, L]$, we define the complex subband frequency error $\mathbf{e}_i^c(\omega)$ as

$$\mathbf{e}_i^c(\omega) = A_{z,i}e^{j\phi_{z,i}} W(\omega - \omega_{z,i}) - \hat{A}_{z,i}e^{j\hat{\phi}_{z,i}} W(\omega - \hat{\omega}_{z,i}), \tag{18}$$

where we define $\hat{A}_{z,i}$, $\hat{\omega}_{z,i}$, and $\hat{\phi}_{z,i}$ respectively as the amplitude, frequency, and phase of the sinusoid that represent the estimated mixture complex spectrum at the ith frequency subband. By setting the estimated mixture phase in (18) equal to the mixture phase sampled at $\omega_{z,i}$ ($\hat{\phi}_{z,i} = \phi_{z,i}$), the absolute error in subbands becomes

$$e_i(\omega) = \left|A_{z,i}W(\omega - \omega_{z,i}) - \hat{A}_{z,i}W(\omega - \hat{\omega}_{z,i})\right|, \quad \omega \in \Omega_i. \tag{19}$$

which has already been used as the MMSE criterion for estimating the sinusoidal parameters [32]. Similar to [34], the summation of the residual error in (18), in fact, approximates the full band spectral distortion given by $\int_{-\pi}^{\pi} |S_z(\omega) - \hat{S}_z(\omega)|^2 d\omega$. Minimization of the residual error at each frequency subband takes advantage of the fact that the error at narrow enough subbands can well be approximated as white noise [39].

To estimate the amplitude and frequency vectors for each of the underlying signals, mixture estimation is performed. Let $\hat{A}_{z,i}^{r,q}, \hat{\omega}_{z,i}^{r,q}$ as the sinusoidal parameters representative taken from the ith frequency band in (14). Using speaker codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$, (14) becomes $f(C(\hat{\boldsymbol{\alpha}}_r, \hat{\boldsymbol{\omega}}_r, \boldsymbol{\phi}_z), C(\hat{\boldsymbol{\alpha}}_q, \hat{\boldsymbol{\omega}}_q, \boldsymbol{\phi}_z), \hat{\rho})$ at each frequency subband i where $\{\hat{\boldsymbol{\alpha}}_r, \hat{\boldsymbol{\omega}}_r\}_{r=1}^M$ and $\{\hat{\boldsymbol{\alpha}}_q, \hat{\boldsymbol{\omega}}_q\}_{q=1}^M$ with $\omega \in \Omega_i$ are the amplitude-frequency codevectors with r

and q as the codebook indices selected from $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively. We further define the mixture estimation difference indicated by $\mathbf{e}_{r,q,i}$ defined for frequency subbands $i \in [1, L]$ as,

$$e_{r,q,i}(\omega) = \left| A_{z,i} W(\omega - \omega_{z,i}) - \hat{A}_{z,i}^{r,q} W\left(\omega - \hat{\omega}_{z,i}^{r,q}\right) \right|, \ \omega \in \Omega_i. \tag{20}$$

Finally, the mixture estimation is carried out by searching for the optimal codevectors (pair of amplitude and frequency) of the codebooks by minimizing

$$J_{r,q} = \sum_{i=1}^{L} e_{r,q,i}^2(\omega), \tag{21}$$

where $\mathbf{e}_{r,q,i}$ is the error vector composed of $e_{r,q,i}(\omega)$ at all frequency subbands with $\omega \in \Omega_i$ and $i \in [1, L]$. We emphasize that the speaker codebooks we use here are in the form of a sinusoidal coder presented in [29], in which each codevector entry is composed of two parts denoted as $(\{\boldsymbol{\alpha}, \boldsymbol{\omega}\})$, sinusoidal amplitude and its corresponding frequencies which determines where the amplitudes are located in the spectrum. To minimize (21), we are required to do search on pairs of codevectors (consisting of amplitudes and frequencies) to determine the optimal pair for signal reconstruction, that is,

$$\{r^*, q^*\} = \arg \min_{\{r,q\} \in \mathbb{C}_1 \times \mathbb{C}_2} J_{r,q}\left(\{\hat{\boldsymbol{\alpha}}_r, \hat{\boldsymbol{\omega}}_r\}, \{\hat{\boldsymbol{\alpha}}_q, \hat{\boldsymbol{\omega}}_q\}\right). \tag{22}$$

We note that the frequency vectors $\hat{\boldsymbol{\omega}}_r$ and $\hat{\boldsymbol{\omega}}_q$ are not the same as frequencies of sinusoids of mixture $\boldsymbol{\omega}_z$, but selected such that they together minimize the cost function in (22). Note that, even after knowing the estimated SSR level and identities of the speakers, exhaustive search of (22) requires $\mathcal{O}(M^2)$ evaluations of the cost function in (22) for *all frames*, which is impractical. Considerable time saving, still retaining high separation quality, can be obtained by using an iterative search as follows. We start with random r, and keep it fixed while optimizing with respect to q, then switching the roles. This requires a total number of $\mathcal{O}(M \times I)$ evaluations of (22), where we particularly set $I = 3$ iterations. This leads to practical speed-up factor of 700:1 for a codebook size $M = 2048$.

F. Signal Reconstruction

The Wiener filter is a classical speech enhancement method that relies on the MMSE estimation to restore the underlying clean signal. Previous studies utilized the Wiener filter [40] operate in the STFT domain. Here we propose to use magnitude ratio filters in the form of square root Wiener filters [40]. According to our preliminary experiments in [20], the reconstruction filters defined in the sinusoidal domain, improve the separation quality as compared to their STFT counterparts. From the definition of the parametric Wiener filter [40] we have:

$$G(\omega) = \left(\frac{P_1(\omega)}{P_1(\omega) + P_2(\omega)} \right)^{\beta} \tag{23}$$

where $P_i(\omega)$ with $i \in \{1, 2\}$ are the power spectra of the signals, which are approximated by the periodograms $|S_i(\omega)|^2$, and the parameter β determines attenuation at different signal-to-noise ratio levels. From the speech enhancement results in [40], it is known that higher values of β result in more attenuation of the interfering signal. However, this achievement comes at the price of increased speech distortion. According to our separation experiments, for signal reconstruction based on the found sinusoidal parameters, throughout our experiments, we use *square root* Wiener filters ($\beta = 0.5$) instead of the conventional Wiener filters ($\beta = 1$).

For synthesizing the separated signals, we produce square root Wiener filters based on sinusoidal feature and apply them to the mixture to recover the unknown signals. Like other separation methods reported in [21], we employ the mixture phase, $\boldsymbol{\phi}_z$ for re-synthesizing the separated outputs. The estimated amplitude-frequency codevectors found in (22) are used to reconstruct their corresponding amplitude spectrum estimates $C(\hat{\boldsymbol{\alpha}}_{r^*}, \hat{\boldsymbol{\omega}}_{r^*}, \boldsymbol{\phi}_z)$ and $C(\hat{\boldsymbol{\alpha}}_{q^*}, \hat{\boldsymbol{\omega}}_{q^*}, \boldsymbol{\phi}_z)$ which are further used to produce square root Wiener filters as below

$$\hat{G}_1(\omega) = \frac{C(\hat{\boldsymbol{\alpha}}_{r^*}, \hat{\boldsymbol{\omega}}_{r^*}, \boldsymbol{\phi}_z)}{\sqrt{C^2(\hat{\boldsymbol{\alpha}}_{r^*}, \hat{\boldsymbol{\omega}}_{r^*}, \boldsymbol{\phi}_z) + C^2(\hat{\boldsymbol{\alpha}}_{q^*}, \hat{\boldsymbol{\omega}}_{q^*}, \boldsymbol{\phi}_z)}}, \tag{24}$$

$$\hat{G}_2(\omega) = \frac{C(\hat{\boldsymbol{\alpha}}_{q^*}, \hat{\boldsymbol{\omega}}_{q^*}, \boldsymbol{\phi}_z)}{\sqrt{C^2(\hat{\boldsymbol{\alpha}}_{r^*}, \hat{\boldsymbol{\omega}}_{r^*}, \boldsymbol{\phi}_z) + C^2(\hat{\boldsymbol{\alpha}}_{q^*}, \hat{\boldsymbol{\omega}}_{q^*}, \boldsymbol{\phi}_z)}}. \tag{25}$$

Accordingly, the separated output time domain signals are given after taking F-point inverse DFT:

$$\hat{s}_1(n) = \mathrm{DFT}_F^{-1}\left\{ \hat{G}_1(\omega) S_z(\omega) e^{j\phi_z(\omega)} \right\} \tag{26}$$

$$\hat{s}_2(n) = \mathrm{DFT}_F^{-1}\left\{ \hat{G}_2(\omega) S_z(\omega) e^{j\phi_z(\omega)} \right\}. \tag{27}$$

III. Results

A. Dataset and System Setup

The proposed speech separation system is evaluated on the speech separation corpus provided in [21]. This corpus consists of 34, 000 distinct utterances from 34 speakers (18 males and 16 females). The sentences follow a command-like structure with a unique grammatical structure as six word commands such as "*bin white at p nine soon*". Each sentence in the database is composed of verb, color, preposition, letter, digit and coda. The keywords emphasized for speech intelligibility or recognition task in challenge are the items in position 2, 4, and 5 referring to color, letter and digit, respectively. The possible choices for color are green, blue, red, and white. The possible letters are 25 English alphabet letters and the digits are selected from 0 to 9.

For each speaker, 500 clean utterances are provided for training purposes. The test data is a mixture of target and masker speakers mixed at six SSR levels ranging from -9 dB to 6 dB. For each of the six test sets, 600 utterances are provided of which 200 are for same gender (SG), 179 for different gender (DG), and 221 for same talker (ST). The sentences were originally sampled at 25 kHz. We decrease the sampling rate to 16 kHz (some additional experiments are also carried out at 8 kHz).

For speaker identification, we extract features from 30 ms Hamming-windowed frames using a frame shift of 15 ms. A 27-channel mel-frequency filterbank is applied on DFT spectrum to extract 12-dimensional MFCCs, followed by appending Δ and Δ^2 coefficients, and using an energy-based voice activity detector for extracting the feature vectors. We add the signals

TABLE II
SPEAKER IDENTIFICATION ACCURACY (% CORRECT) WHERE BOTH SPEAKERS ARE CORRECTLY FOUND

SSR (dB)	-9	-6	-3	0	3	6	Average
Iroquois [8]	96.5	98.1	98.2	99.0	99.1	98.4	98.2
Saeidi *et. al.* [18]	86.7	93.0	97.1	96.2	92.8	91.6	92.9
Proposed	87.5	93.2	97.2	96.2	92.9	91.7	93.2

with an average frame-level SSR to construct the universal background model (UBM) and the target speaker GMMs. For each of the 34 target speakers, 50 randomly chosen files from each speaker are mixed at SSR levels $g \in \{-9, -6, -3, 0, 3, 6\}$ dB with 50 random files from all other speakers, which gives us about 180 hours of speech for UBM training. The number of Gaussians is set to $M = 2048$.

Each SSR-dependent GMM, λ_{ig}, is trained by mixing 100 random files from the ith speaker with 100 random files from all other speakers which gives about 1.8 hours data for training. The relevance factors in MAP adaptation were set to $\beta = 16$ for training the speaker models and $\beta = 0$ for training the test utterance models, respectively. The choice of $\beta = 0$ for the test utterance was done due to short length of data for adaptation.

Table II shows the accuracy of the proposed speaker identification module for finding both target and masker speakers. An average accuracy of 93.2% is achieved using the proposed SID module. Considering D speakers, M Gaussians and G SSR-levels, the number of Gaussian evaluations for the speaker recognition system are $\mathcal{O}(DM^D)$ for the *Iroquois* system [8]. The proposed approach, on the other hand, has computational complexity of $\mathcal{O}(DGM)$ only. Therefore, the proposed SID module is much faster in operation in exchange of reduced accuracy.

For separation, we extract features by employing a Hann window of length 32 ms and shift of 8 ms. We use split-VQ based on sinusoidal parameters [29]. The source models are divided into magnitude spectrum and frequency parts where each entry is composed of a sinusoidal amplitude vector and several sinusoidal frequency vectors as its candidates. According to previous experiments, we set the sinusoidal model order to $L = 100$ for 16 kHz and $L = 50$ for 8 kHz [9]. For speaker modeling, we use 11 bits for amplitude and 3 bits for frequency part in the sinusoidal coder. This results in codebook size of 2048 in split-VQ for modeling sinusoidal features. Studying the other features effect in the subsequent subsections, the same codebook size of 2048 is also used for speakers' VQ models. The pre-trained speaker codebooks are then used in the test phase to guide the speech separation. The codebooks are used for both the mixture estimator and the double-talk detector (Fig. 2). For the mixture estimator given in (14), we used the first 5 terms of the elliptic series in (15).

As our benchmark methods, we use the two systems in [7] and [8] participated in the SCSS challenge. We report the separation results on the outputs obtained by the *super-human* speech recognition system [8] as top-performing separation systems in the challenge. This system even outperforms human listeners in some of the speech recognition tasks [21]. As the second benchmark system, we use another top-performing separation system, "speaker-adapted full system" proposed in [7] (see Table II in [7]), where *Iroquois* [8] system was used for estimating the speaker identity and the SSR level both in [7] and [8].

We had access only to a limited number of separated clips[1] for the system in [8], where the authors in [7] supplied their separated signals on the whole GRID corpus. To this end, we evaluate the performance of the proposed system in terms of four experiments:

- Demonstrating how the mixture estimation is performed in sinusoidal domain using the proposed MMSE mixture amplitude estimator in (14) and studying its impact on performance compared to the full band STFT case.
- Subcomponent comparison versus the existing state-of-the-art.
- Comparing the proposed method versus benchmark in [7] employing the whole corpus using perceptual evaluation of speech quality scores (PESQ) and short-time objective intelligibility measure (STOI).
- Comparing proposed method versus benchmarks in [7] and [8] on limited number of clips using different objective and subjective measures.

B. Experiment 1: Case Study for MMSE Mixture Amplitude Estimator in Sinusoid

We select the mixture of two female speakers 7 and 11 from GRID corpus test set mixed at $\mathrm{SSR} = 0$ dB. We represent speech signals using limited number of sinusoids where frequencies and amplitudes are obtained using the peak picking on the mel-scale as described in (6–7). We consider two scenarios: i) ideal case, where the speaker spectra are known, and ii) estimated by the optimal codebook entry, determined as the result of the codebook search in (22). The results for the ideal scenario and estimated from codebook are shown in Fig. 4 on the right and the left panels, respectively. Fig. 4 (right) shows how the proposed sinusoidal MMSE mixture amplitude estimator works by minimizing the error over the harmonic lobes of the sinusoids, estimated per frequency subbands, defined in (20). Subplot (a) shows the observed mixture spectrum of two speakers and the mixture estimated using the proposed MMSE estimator in (14). Subplot (b) displays the mixture estimation error power in decibels for both STFT and sinusoidal features. Subplots (c) and (d) illustrate the original spectra of the two underlying speakers, as well as the STFT and sinusoidal spectrum amplitude. Comparing the MSE results of full-band and sinusoidal shown in subplot (b), it is concluded that the proposed sinusoidal MMSE amplitude estimator defined in (18) well approximates the full-band mixture estimation error defined in (17). For visual clarity, we use dB-scale for the spectral magnitudes but all computations use the original spectral magnitude values. We have only shown the frequencies in the range of [0, 3800] Hz at a sampling frequency of 8 kHz.

As a second scenario, we compare the results of mixture estimation in full-band STFT domain and sinusoidal features by performing codebook search on the STFT codebooks and sinusoidal split-VQ codebooks, respectively. The results are shown

[1]The clips are Clip 1: target sp6:bwba masker sp30:pgah6a (mixed at -3 dB), Clip 2: target sp14:lwax8s masker sp22:bgwf7n (mixed at 0 dB), Clip 3: target sp33:bwid1a masker sp33:lgii3s (mixed at -6 dB) and Clip 4: target sp5:swah6n masker sp5:bbir4p (mixed at 0 dB) signal-to-signal ratio.

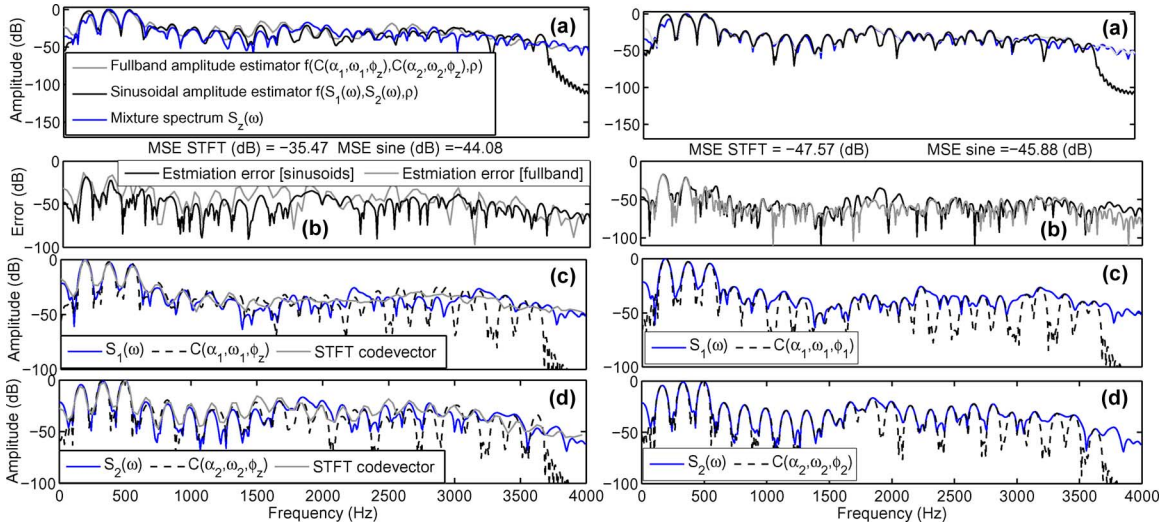

Fig. 4. Shown are the magnitude spectrum for (left) codebook search scenario (right) ideal scenario. The descriptions of each panel: (a) shows the original and estimated mixture spectrum amplitude denoted by $S_z(\omega)$ and $f(C(\boldsymbol{\alpha}_1,\boldsymbol{\omega}_1,\boldsymbol{\phi}_z),C(\boldsymbol{\alpha}_2,\boldsymbol{\omega}_2,\boldsymbol{\phi}_z),\rho)$, respectively, (b) mixture estimation error power $e(\omega)$ in decibels. The MSE value for full-band and sinusoidal cases are reported for the bottom plot, and (c) speaker one: $S_1(\omega)$ and $C(\boldsymbol{\alpha}_1,\boldsymbol{\omega}_1,\boldsymbol{\phi}_z)$, (d) speaker two: $S_2(\omega)$ and $C(\boldsymbol{\alpha}_2,\boldsymbol{\omega}_2,\boldsymbol{\phi}_z)$.

in the left panel of Fig. 4. The sinusoidal MMSE amplitude estimator achieves a lower MSE compared to the STFT case. The selected codevectors result also in a more accurate amplitude spectrum representation than the STFT scenario (see subplots (c) and (d) in the left panel).

C. Experiment 2: Analysis of the System Sub-Components: Features, Frequency Warping, Mixture Estimator, Type of Mask

In this subsection, we experimentally compare the choice of each component in our full system to alternative state-of-the-art components. To this end, we evaluate the separation performance in terms of different attributes: i) joint feature and mixture estimator, ii) feature selection independent of speaker model, iii) quantization effect, and iv) different filters for signal reconstruction. As our experiment setup, we selected two speakers, 9 and 19, from the GRID corpus for mixing. As our quality assessment measure, we chose PESQ and the results are averaged over 50 utterances.

The following alternatives for the feature and mixture estimator are considered:

- Features: Gammatone auditory scale filter bank (GTFB), mel-frequency band energy (MFBE), STFT and sinusoidal feature. For GTFB features, we considered 128 log-energy of gammatone auditory scale filter-bank whose filters are quasi-logarithmatically spaced, based on the equivalent rectangular bandwidth (ERB)-scale [2]. The bandwidth increases with center frequency from about 35 Hz at 100 Hz to around 670 Hz at 6000 Hz. We select MFBE features as a commonly used auditory scale features in variety of applications. Following the setup in [30], to extract MFBE features, we designed the filterbank in ERB scale and applied the filterbank to the power spectrum of signal. In the reconstruction stage a pseudo-inverse of the filterbank is utilized to minimize the Euclidean norm, as suggested in [41]. The number of filterbanks was set to 60 based on our preliminary experiments.
- Mixture estimator: MMSE in log-power spectrum, power spectrum, spectrum amplitude domain (proposed), sinusoidal estimator of [9], and subband perceptually weighted transformation (SPWT). SPWT uses STFT features and employs a perceptually weighted spectral distortion in frequency subbands by imposing a weighting to emphasize different frequency division in an uneven manner in contrast to STFT case [42]. We used four frequency subbands division in Mel-scale as it led to the highest PESQ as reported in [42].

In the proposed system, the codebook indices r and q are jointly estimated from the mixed signal using (22). In turn, if we estimate r and q (the codevector indices in the two codebooks) from the original spectra, $S_1(\omega)$ and $S_2(\omega)$, using

$$r^* = \arg\min_{\mathbf{c}_r\in\mathbb{C}_1} \|C(\boldsymbol{\alpha}_1,\boldsymbol{\omega}_1,\boldsymbol{\phi}_1) - C(\boldsymbol{\alpha}_r,\boldsymbol{\omega}_r,\boldsymbol{\phi}_1)\|_2^2, \quad (28)$$

$$q^* = \arg\min_{\mathbf{c}_q\in\mathbb{C}_2} \|C(\boldsymbol{\alpha}_2,\boldsymbol{\omega}_2,\boldsymbol{\phi}_2) - C(\boldsymbol{\alpha}_q,\boldsymbol{\omega}_q,\boldsymbol{\phi}_2)\|\|_2^2, \quad (29)$$

where $\boldsymbol{\alpha}_k,\boldsymbol{\omega}_k$ are the amplitude-frequency feature set obtained by applying sinusoidal feature extraction (5–6) on the kth speaker signal, $s_k(n)$, we call the set-up as VQ-based upper bound. The VQ-based upper bound is the best possible performance obtainable by the proposed model-driven speech separation approach [43].

1) Studying the Joint Impact of Feature and Mixture Estimator: Here we evaluate the separation performance in terms of two attributes (1) feature domain representation and (2) mixture estimator selection. To this end, we select STFT, mel-frequency band energy (MFBE), sinusoidal feature space while the mixture estimators are MMSE in log-power spectrum, power spectrum, and spectrum amplitude domain (proposed). In addition, to locate the performance of the proposed algorithm among the previously similar ones, we also report the results obtained by ML sinusoidal estimator [9] and SPWT [42]. The separation performance results are shown in Fig. 5. We make the following observations:

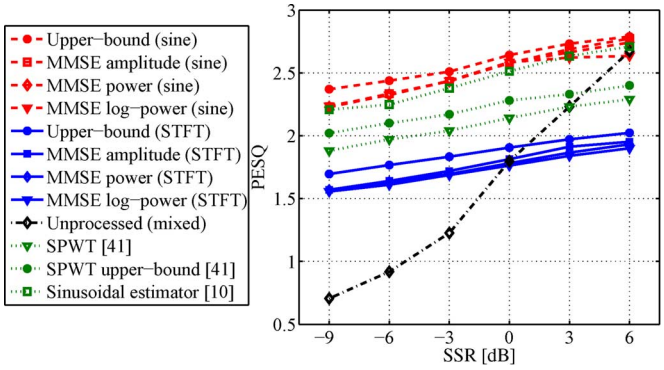

Fig. 5. Comparing the separation performance of the proposed system in terms of PESQ for different combination of mixture estimators MMSE estimator in log-power, power, and amplitude domain (proposed) with different features (STFT and sinusoidal). The performance of subband perceptually weighted transform in [42] and sinusoidal estimator in [9] are also included.

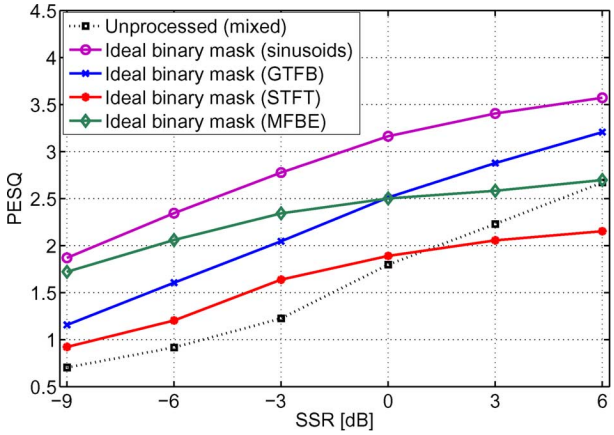

Fig. 6. Studying the feature impact independent of the quantization. Showing the separation performance obtained by using ideal binary mask for different features.

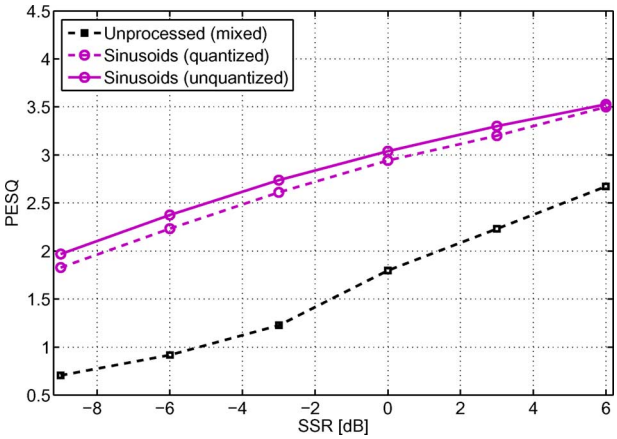

Fig. 7. Showing the quantization effect on sinusoidal features in an oracle separation scenario.

- For a given speaker codebook, the closer the method asymptotes to its VQ-based upper-bound performance, the more accurate the mixture estimator is. we observe that the differences between STFT- and sinusoidal-based estimators are not significant.
- The impact of replacing STFT features with sinusoidal features is observed by comparing the VQ-based upper-bound performance obtained by the selected features in Fig. 5. It is observed that sinusoidal features offer a considerably higher upper-bound compared to the STFT.
- For the STFT features, the proposed MMSE amplitude estimator results in improved separation performance compared to both the MMSE log-power and the MMSE power estimators for $\mathrm{SSR} > 0$ decibels. For $\mathrm{SSR} < 0$ all the MMSE estimators achieve similar performance. The same trend is also observed for the sinusoidal features. In particular, when SSR increases, the performance of the amplitude MMSE estimator approaches the VQ-based upper-bound performance.
- The proposed MMSE amplitude estimator in sinusoid achieves slightly better performance compared to the sinusoidal estimator presented in [9] and SPWT.

From the PESQ results shown in Fig. 5, we conclude that the impact of the selected feature is more pronounced than that of different mixture estimators.

2) Studying the Impact of the Selected Feature Independent of the Speaker Codebook: To assess the separation results for different features without considering the effect of model type (VQ) and its selected order, we present the separation results for ideal binary mask (IBM) for different features. The ideal binary mask is defined as the mask produced by keeping all time-frequency cells where the target speaker dominates the interfering one and removing those where the target is masked by the interfering speaker [2]. The results are shown in Fig. 6. It is concluded that replacing STFT with auditory transform or sinusoidal, improves the signal quality results across all SSRs.

3) Studying the Effect of Quantization: In model-based speech separation, it is required to capture speaker characteristics with a model. However, as in any modeling technique, the quantization process in representing an actual speech event with an average model, degrades the achievable separation performance. The impact of the quantization step on the separation performance is evaluated throughout the experiments by reporting the "VQ-based upper-bound" performance shown in Fig. 7. In an oracle separation scenario, we conduct an experiment to study the effect of replacing the quantized sinusoidal features in (24–25) with the unquantized features. For quantized sinusoids we use $C(\boldsymbol{\alpha}_r, \boldsymbol{\omega}_r, \boldsymbol{\phi}_1)$ and $C(\boldsymbol{\alpha}_q, \boldsymbol{\omega}_q, \boldsymbol{\phi}_2)$ with $r \in \mathbb{C}_1$ and $q \in \mathbb{C}_2$ while for the unquantized features, we directly use $C(\boldsymbol{\alpha}_1, \boldsymbol{\omega}_1, \boldsymbol{\phi}_1)$ and $C(\boldsymbol{\alpha}_2, \boldsymbol{\omega}_2, \boldsymbol{\phi}_2)$. This experiment demonstrates how accurately the quantized sinusoidal features represent the original sinusoidal parameterization. The small gap in PESQ between quantized and unquantized sinusoidal features indicates that the employed split-VQ model represents the sinusoidal parameters of signal accurately. The reason why PESQ scores are increasing as the SSR evolves is that the mixture information is utilized when reconstructing the output signals.

4) Studying the Impact of Different Filters for Signal Reconstruction: First, we compare the two mask methods as follows: i) employing the square root Wiener filters $\hat{G}_1(\omega)$ and $\hat{G}_2(\omega)$ as defined in (24–25); ii) replacing the phase integrated out mixture estimate $\hat{S}_z(\omega)$ in (14), to the denominator of the square root Wiener gain function in (24–25). To recover the corresponding source estimates, each filter is then applied to the mixed signal.

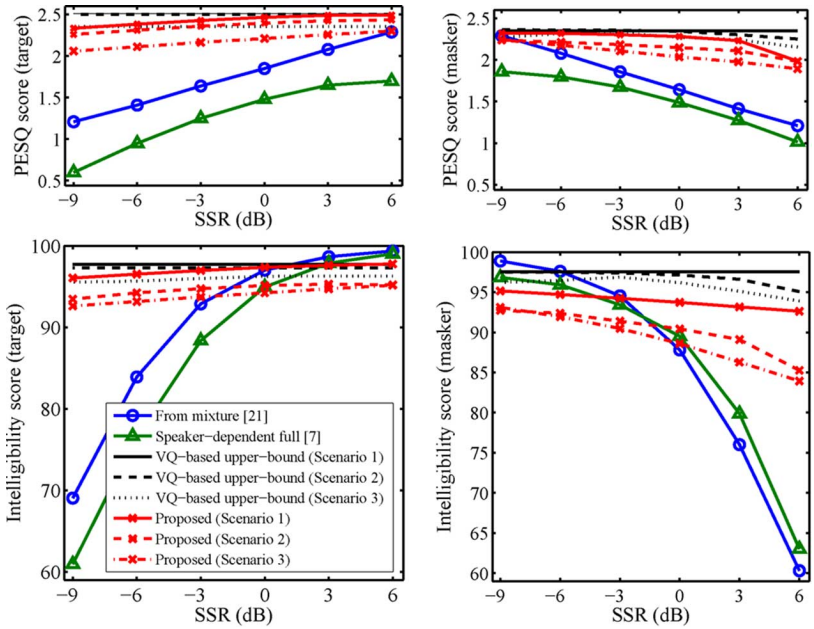

Fig. 8. (Top) perceptual evaluation of speech quality scores (PESQ), and (Bottom) short-time objective intelligibility measure (STOI) scores for target and masker. According to [44], for normal subjective test material the PESQ values lie between 1.0 (bad) and 4.5 (no distortion). According to [46], the intelligibility score lies between 0 (bad) and 100 (no distortion). All the results are reported on the speech separation challenge test data provided in [21].

TABLE III
COMPARING THE SEPARATION PERFORMANCE IN PESQ FOR DIFFERENT CHOICES OF MASK FUNCTION IN SIGNAL RECONSTRUCTION STAGE

Source number	Speaker 1		Speaker 2	
Feature	STFT	Sinusoidal	STFT	Sinusoidal
Filter using (14)	1.77±0.09	2.36±0.09	1.76±0.14	2.11±0.13
Filter using (24-25)	**2.01±0.09**	**2.66±0.10**	**1.91±0.20**	**2.42±0.17**

The two filters differ only in terms of their denominator; To recover the corresponding source estimates, each gain function is then applied to the mixed signal. The results are summarized in Table III. The results obtained for both the STFT and sinusoidal features indicate that improvement is achieved with square root Wiener gain functions ($\hat{G}_1(\omega), \hat{G}_2(\omega)$ in (24–25)) compared to masks with phase integrated out ($\hat{S}_z(\omega)$ in (14)). This is justified from the improvement of 0.3 in PESQ for both speakers.

D. Experiment 3: PESQ & STOI Evaluation on Whole Test Set

To study the performance of the proposed speech separation system, we consider six different setups, covering cases from all parameters known to all parameters estimated. These six setups are shown in the legend of Fig. 8 as scenarios 1, 2 and 3 with their corresponding upper-bounds (which we call *known codebook index*). Parameters that we consider include codebook index, speaker identity and SSR level. The scenarios are defined as:

- Scenario 1: known SID and SSR,
- Scenario 2: estimated SID and known SSR,
- Scenario 3: estimated SID and SSR.

In scenario 1, given the correct SID and SSR level, we investigate the accuracy of the mixture estimation stage. Additionally, we also consider degradations caused by erroneous speaker identities and SSR estimation as in scenarios 2 and 3, respectively.

For objective measurement, we use PESQ [44] as it correlates well with subjective listening scores [45] and STOI [46] since it showed higher correlation with speech intelligibility compared to other existing objective intelligibility models. Fig. 8 shows the separation results in terms of PESQ and STOI obtained for different scenarios. The results obtained from mixture and scores calculated for the separated wave files of [7] are also shown for comparative purposes.

Fig. 8 suggests that the proposed method improves the quality of the separated signals compared to the mixture. According to the masking theorem [47], at low SSR levels, *energetic masking* occurs and the separation system successfully performs in compensating this effect by separating the underlying speakers for each frame. At high SSR levels, *informational masking* is more dominant and the mixed signal itself is more intelligible than the separated signals obtained by separation module. The mixed signal itself achieves higher intelligibility score compared to the separated target signal since the target speaker becomes more dominant. At high SSR levels, the proposed method asymptotically reaches the best possible performance denoted by VQ-based upper bound performance.

The proposed method outperforms the method in [7] in terms of PESQ at all SSR levels. It also improves the intelligibility of the target speaker significantly at low SSR levels (lower than -3 dB). However, the speaker-adapted full system in [7] achieves slightly higher intelligibility scores. By comparing the results of the known (scenario 1) and the estimated speaker identities (scenario 2), the results are generally close to each other. The same conclusion holds also for the *known and estimated SSR levels*. This confirms that the SID and SSR estimates were relatively accurate as suggested by Table II.

Studying different scenarios, the proposed system performs better for *different gender* compared to the *same gender*. A similar observation was reported in [7]. This can be explained by

TABLE IV

THREE SYSTEM COMPARISON WITH DIFFERENT METRICS ON FOUR CLIPS FROM GRID CORPUS. SYSTEMS ARE S1: HERSHEY [8], S2: WEISS [7] AND S3: PROPOSED AT 16 kHz. METRICS ARE STOI [46], CROSS-TALK [48], PESQ [44], SIR [49], SAR [49], SDR [49], $\mathrm{SNR}_{\mathrm{loss}}$ MEASURES [50], QUALITY SCORES [51]. EACH CLIP IS CHARACTERIZED BY ITS MIXING SSR LEVEL AND THE MIXING SCENARIO: DIFFERENT GENDER (DG), SAME GENDER (SG) AND SAME TALKER (ST). IN EACH SUB-COLUMN, THE BEST RESULT IS HIGHLIGHTED WITH **Shaded Bold** FONT

		Target												Masker										p-value	
		Clip 1 (SG -3dB)			Clip 2 (DG 0dB)			Clip 3 (ST -6dB)			Clip 4 (ST 0dB)			Clip 1 (SG -3dB)			Clip 2 (DG 0dB)			Clip 3 (ST -6dB)			Clip 4 (ST 0dB)			S3 vs.	S3 vs.
	Criterion	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2
	STOI	0.77	0.70	**0.79**	0.80	0.82	**0.83**	**0.74**	0.51	**0.74**	**0.85**	0.48	0.75	0.83	0.83	**0.84**	**0.82**	0.70	0.74	0.21	0.43	**0.68**	0.13	0.49	**0.65**	<0.05	<0.05
	Cross-talk	11.5	13.3	**8.0**	10.3	11.9	**5.8**	4.3	**2.3**	5.9	**4.4**	9.4	6.3	10.1	**5.9**	10.4	**10.3**	10.5	**10.3**	13.8	17.3	**11.1**	12.4	**10.1**	**10.1**	>0.05	>0.05
	PESQ	**2.4**	1.4	**2.4**	1.5	1.7	**2.4**	**2.2**	1.7	**2.2**	**2.8**	1.0	2.5	1.4	1.3	**2.0**	**2.2**	1.3	**2.2**	**2.9**	1.0	**2.9**	**2.4**	1.7	**2.4**	>0.05	<0.05
	$\mathrm{SNR}_{\mathrm{loss}}$	0.96	0.98	**0.91**	0.99	0.89	**0.83**	**0.92**	0.98	**0.92**	**0.91**	0.98	**0.91**	0.93	0.99	**0.92**	0.97	0.98	**0.89**	**0.96**	**0.96**	**0.96**	0.97	0.95	**0.90**	<0.05	<0.05
BSS EVAL	SIR	10.8	**20.0**	15.0	0.1	12.6	**16.5**	2.0	2.7	**14.7**	8.4	17.0	**17.4**	2.4	7.8	**15.3**	11.7	14.9	**20.6**	**13.6**	6.9	**13.6**	8.8	-8.7	**16.7**	<0.05	<0.05
BSS EVAL	SAR	**7.9**	1.4	3.9	**43.2**	-1.3	0.4	**6.9**	1.4	1.9	**9.5**	-3.3	2.6	**7.7**	-5.1	-0.8	**9.2**	-0.3	1.8	**12.0**	-6.6	5.6	**8.3**	-6.4	-0.9	<0.05	<0.05
BSS EVAL	SDR	**5.8**	0.1	1.1	**1.0**	0.0	0.2	**2.6**	0.2	0.2	**6.1**	0.1	0.8	**2.8**	0.1	0.2	**8.0**	0.1	1.0	**8.7**	0.0	1.8	**4.2**	0.0	0.3	<0.05	<0.05
PEASS	OPS	**53**	50	36	**42**	19	33	**57**	26	41	**53**	24	30	**45**	23	36	**66**	25	52	**73**	18	41	**64**	33	34	<0.05	<0.05
PEASS	TPS	**82**	69	77	59	79	**86**	73	26	**75**	**80**	26	73	**76**	53	59	**72**	50	**72**	**78**	32	75	75	32	**76**	>0.05	<0.05
PEASS	IPS	19	83	**84**	79	77	**80**	72	75	**82**	75	66	**78**	69	77	**81**	65	71	**79**	70	71	**78**	78	**80**	76	<0.05	<0.05
PEASS	APS	**60**	37	36	**90**	14	26	**63**	16	33	**58**	13	20	**61**	11	31	**72**	14	43	**76**	9.0	36	**63**	18	29	<0.05	<0.05

the different time-frequency masking patterns and physiological differences in the vocal characteristics of male and female speakers. Thus, the underlying sources are less overlapped compared to other scenarios.

E. Experiment 4: Performance Evaluation on a Subset of Test Data

In the following, we compare the proposed method to those proposed in [7], [8] for selected clips from test dataset composed of same gender, different gender and same talker scenarios. The separation results are summarized in Table IV. For each of the measures in this experiment, the significance level for each paired t-test (p-value) is shown in the last column in Table IV. The p-values determine whether the results obtained by the proposed method are significantly different than benchmark methods. The following observations are made:

1) STOI [46]: The proposed method achieves better performance compared to the baseline methods.

2) Cross-Talk [48]: An ideal separation system would filter out any trace of the interfering speaker signal in the mixture. As a proof of concept, we use the amount of *cross-talk* [48] remaining in the separated output signal for comparing different separation methods. From the cross-talk scores, we conclude that the proposed SCSS method often introduces less cross-talk compared to [7]. Although the differences are not statistically significant, we observe that the proposed system leads to relatively less or comparable amount of cross-talk in most of the cases compared to [7] and [8], respectively.

3) PESQ [40]: The proposed system yields improved results over the method in [7].

4) SNR_{loss} [50]: This measure was found appropriate in predicting speech intelligibility in different noisy conditions, in the sense of producing a higher correlation for predicting sentence recognition in noisy conditions ($r = -0.82$ higher than $r = 0.77$ for PESQ). From the $\mathrm{SNR}_{\mathrm{loss}}$ results we observe that the proposed method consistently outperforms the competitive methods.

5) BSS EVAL Metrics [49]: To enable comparison with other source separation algorithms, we evaluate the separation results in terms of the metrics proposed in blind source separation evaluation toolkit (BSS EVAL) [49]. The following observations are made:

- The proposed method achieves a better signal-to-interference ratio (SIR) performance compared to both benchmark methods. This improvement in SIR compared to [8] is attained at the price of introducing more artifacts, i.e. producing lower signal-to-artifact ratio (SAR). This implies that a separation quality with less cross-talk is feasible but introduces more artifacts. This is analogous to the tradeoff between speech distortion minimization and cross-talk suppression provided by the square root Wiener filter based on sinusoids discussed in Section III-F. This suggests that the proposed method is often better at rejecting interference when recovering the target speaker. Similar trade-off between SIR and SAR result was independently reported in [49].
- The proposed method achieves better SAR and SDR scores compared to [7] but lower than [8] which achieves the highest SDR and SAR scores. The signal-to-distortion (SDR) measure takes into account both interference and noise level in the excerpts and, consequently, has no preference over interference signal or noise power; therefore, the same level of each will degrade the SDR metric by the same amount.

6) PEASS [51]: We report the separation results in terms of the state-of-the-art objective metrics called *perceptual evaluation methods for audio source separation* (PEASS) adopted for the 2010 signal separation evaluation campaign (SiSEC) [51]. We use the four quality scores proposed in PEASS toolkit [51]: overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS). OPS measures how close the separated signal is to the clean signal, TPS measures how close the target-related part of the enhanced signal is to the clean reference signal, IPS measures the interference cancellation in the separated signal, and finally, APS shows how close the enhanced signal is to the clean one in terms of having no artifacts. We make the following observations:

- The APS results are in line with SAR results confirming that [8] produces least artifacts. This might be because [8] employs both dynamic speaker models and grammar constraints. Meanwhile, the proposed method attains higher SAR and APS performance compared to [7].
- According to TPS results, both [8] and the proposed method achieve higher performance compared to [7]. The

paired test outcome between the TPS scores of [8] and the proposed method indicates insignificant difference.

- The system in [8] achieves the highest OPS scores among the three systems. The proposed system achieves higher performance compared to [7].
- The outcome of paired tests on IPS scores confirms those obtained on SIR, indicating statistically significant difference of the proposed method over others.

7) MUSHRA [52]: To assess the perceived quality obtained by the different separation methods, as our first subjective measure, we conduct subjective test using the so-called MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test as described in [52]. The MUSHRA test is a double blind test for the subjective assessment of intermediate quality level benefits obtained by different methods (via displaying all stimuli at the same time). The MUSHRA test enables simultaneous comparison of different separation methods directly.

We conducted the listening experiments in a silent room using high quality audio device with firewire interface for digital-to-analog conversion and AKG K240 MKII headphones. To ease the test procedure, we prepared a graphical user interface (GUI) in MATLAB. Seven untrained listeners participated in the test (none of the authors were included). The excerpts consisted of the hidden reference (HR) showing the known quality on the scale; it is used to check the consistency of the responses of a subject. A high score is expected for HR. We also include the mixed signal (without any separation) as an anchor point to enable comparison of separated signal and mixture qualities. This reflects how hard it was to perceive the reference signal when listening to the mixture. The remaining four excerpts are the separated signals obtained by *super-human* speech recognition system [8], speaker-adapted full system [7], and our proposed method configured for both 8 kHz and 16 kHz sampling frequencies. The excerpts were randomly chosen and played for each subject. The excerpts used in subjective tests are downloadable from the webpage: http://www.audis-itn.eu/wiki/Demopage2. The listeners were asked to rank eight separated signals relative to a known reference on a scale of 0 to 100.

The MUSHRA test results are reported in terms of the mean opinion score (MOS) and 95% confidence intervals [53] calculated according to the standard as described in ITU-R BS.1534-1 [52]. Fig. 9 shows the mean opinion score (MOS) for comparing the separation results obtained by different methods discussed in this paper. We observe that the maximum and minimum scores were obtained at hidden reference and speech mixture, respectively, as expected. Furthermore, the proposed method at 16 kHz achieves better performance compared to [7]. The difference between the performance of the method studied in [8] and the proposed one is not statistically significant. This result confirms the PESQ score observation. The proposed method at 8 kHz also achieves comparable result with [8] and [7].

8) Speech Intelligibility [54]: Following the principle and standard described in [54], as our second subjective measurement, we conducted a test to assess speech intelligibility of the separated signals. We chose seven listeners (different from those that participated in the MUSHRA test) and eight segments to be played for each listener. We asked the listeners to identify

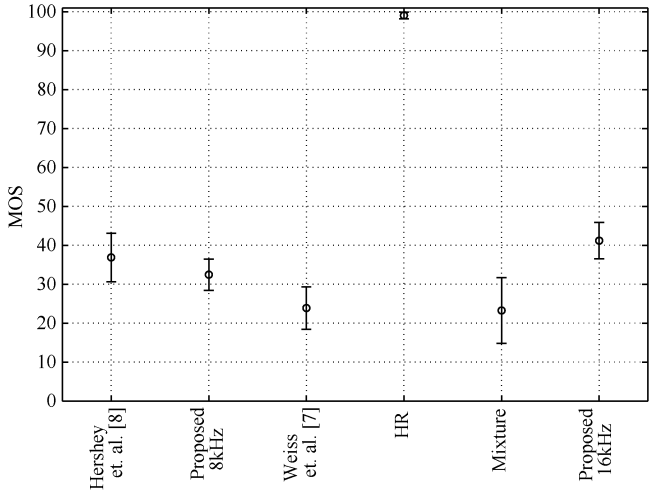

Fig. 9. Results of the MUSHRA listening test for different separation methods averaged over all excerpts and listeners. Error bars indicate 95% confidence intervals.

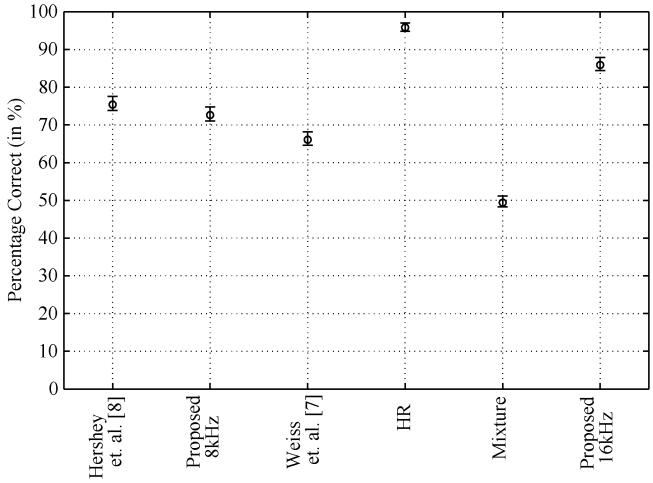

Fig. 10. Speech intelligibility test results. The calculated percentage of correct keywords is averaged over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

color, letter, and digit spoken during each of the played segments. The listeners were required to enter their results using a GUI in MATLAB, which enabled listeners to enter their results both accurately and comfortably. On average, it took 15 minutes per listener to complete the test.

Fig. 10 shows the results of the intelligibility test averaged over all excerpts and listeners. We observe that the proposed method at 16 kHz achieves higher speech intelligibility compared to the methods in [7] and [8]. This result is in agreement with our observations on both $\mathrm{SNR}_{\mathrm{loss}}$ and STOI. The mixed signal also has the lowest score while the hidden reference signal achieves the highest intelligibility score, as expected.

9) ASR Results: Finally, we also configured an automatic speech recognition system using mean subtraction, variance normalization, and ARMA filtering (MVA) [55], which gave an overall recognition accuracy of 52.3% [56]. Comparing the result with those of the systems reported by the other participants in the separation challenge [21, Table 1], we observed that our system ranks on the range of median out of all methods; located below [8] 78.4% but above [7] 48.0%.

IV. Discussion

Both the objective and subjective results show that fairly good separation quality and high interference rejection capability were achieved, in comparison to other methods in the field. In particular, the subjective measurements indicate that the proposed system improves both quality and intelligibility of the signal and achieves a performance comparable to the systems in [7] and [8]. Although the performance of proposed system in light of speaker identification and automatic speech recognition is not better than the top-performing systems but it is comparable with other algorithms in speech separation challenge [21]. Our proposed separation system separates the mixture frame-by-frame and is appropriate for low-delay applications, such as speech coding.

The proposed system, like other current separation systems, still has some limitations. The training samples used to train the speaker models are noise-free and relatively long and the evaluation corpus consists of only digitally added mixtures. Additionally, the gains of the underlying speakers in the mixture are assumed to be constant and we have a mixture of two speakers only. We also neglected the environmental or background noise effects, as well as the reverberation problem. In practice, each one of these issues and their effect on the overall separation performance should be carefully studied. Future work should systematically address how these simplifying, yet restrictive and impractical pre-assumptions could be relaxed. As an example, [57] provides a new corpus for noise-robust speech processing research where the goal is to prepare realistic and natural reverberant environments using many simultaneous sound sources.

The improvement using the proposed MMSE sinusoidal mixture estimator over our previous sinusoidal mixture estimator can be elaborated as follows. The ML sinusoidal mixture estimator presented in [9] ignores the cross-term components between the underlying speakers' spectra at each frame, as well as their phase differences. In some situations, the interference sinusoidal components, play a critical role and can change the position of spectral peaks completely. The proposed sinusoidal MMSE estimator presented in this work, in turn, considers the cross terms and integrates out the phase difference based on uniformity assumption of the speech phase. This explains why the MMSE sinusoidal mixture estimator achieves improved MSE compared to the sinusoidal mixture estimator of [9]. Finally, similar to other sinusoidal modeling systems like [31], the proposed method introduces some buzziness for unvoiced segments. As a future work and room for improving the performance, it is possible to consider more complex modeling for speech and jointly estimating voicing states and sinusoidal model parameters of the underlying signals.

The presented system showed high perceived quality and intelligibility of the separated signals. The results obtained in the speech intelligibility test can be interpreted as the human speech recognition results obtained from the separated signals. There are two possible reasons why the ASR results are in disagreement with our signal quality scores. Firstly, the word error rate metric of ASR does not correlate with those used for assessing the signal quality. Secondly, evaluating the separation performance using ASR systems depends on the speech recognizer configuration, features, training of acoustic and language models. It is not trivial to configure an ASR-system optimized for STFT-like features, to work well on sinusoidally coded speech. Therefore, improvement of the automatic speech recognition performance of the proposed system is left as a future work.

V. Conclusion

We presented a novel joint speaker identification and speech separation system for solving the single-channel speech separation problem. For the separation part, we proposed a double-talk/single-talk detector followed by a minimum mean square error mixture estimator for mixture magnitude spectrum operating in the sinusoidal domain. Importantly, the proposed method does not require pitch estimates and is based on sinusoidal parameters. We relaxed the *a priori* knowledge of speaker identities and the underlying signal-to-signal ratio (SSR) levels in the mixture by proposing a novel speaker identification and SSR estimation method. The proposed system was evaluated on the test dataset provided in the *speech separation challenge*. Compared to previous studies that mostly report speech recognition accuracies, additionally, we focused on reporting the signal quality performance obtained by different separation methods. From the experimental results of various objective and subjective measurements, we conclude that the proposed method improves the signal quality and the intelligibility of the separated signals compared to the mixture and the tested state-of-the-art methods, while it does not meet the performance of state-of-the-art systems in terms of speaker identification and automatic speech recognition accuracy. In many cases, the method offered separated signals with less cross-talk via a high interference rejection capability. Considering different objective and subjective metrics, evaluated on three systems outputs, we conclude that no single system can produce an output satisfying all the evaluation metrics. By comparing the subjective results with those obtained by objective metrics and performing statistical significance analysis, we conclude that the ranking of the systems changes according to the chosen objective metric. The difference between our objective and subjective results, reveals a mismatch between the performance evaluation in the back end and the parameter estimation stage in the separation stage, when the separation system is used as a pre-processor for a target application, e.g., automatic speech recognition.

Acknowledgment

Authors would like to thank Dr. Ron Weiss and Prof. Dan Ellis for their assistance in sharing their data and Dr. Emmanuel Vincent for his helpful discussions concerning the BSS EVAL and PEASS implementation. We would like to thank Prof. Deliang Wang for his help in implementing and evaluating gammatone filterbank features and Dr. Jon Barker for his helpful discussion concerning the speech intelligibility test. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions which were helpful in improving the paper.

REFERENCES

[1] P. Mowlaee, "New Stategies for Single-Channel Speech Separation," Ph.D. thesis, Institut for Elektroniske Systemer, Aalborg Universitet, Aalborg, Denmark, 2010.

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE Press, 2006.

[3] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[6] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 67–76, 2010.

[7] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 16–29, 2010.

[8] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[9] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1265–1277, Jul. 2011.

[10] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the em algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1993, vol. 2, pp. 728–731.

[11] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 94–111, 2010.

[12] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 5, pp. 957–960.

[13] S. T. Roweis, "One microphone source separation," in *Adv. in Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 793–799.

[14] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003, pp. 1009–1012.

[15] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 66–80, Nov. 2010.

[16] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[17] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4430–4433.

[18] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2010, pp. 4545–4548.

[19] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," in *Proc. IEEE Int. Symp. Telecomm.*, Aug. 2008, pp. 543–547.

[20] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4262–4266.

[21] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[22] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[23] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Co-channel speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 828–831.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Elsevier Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

[25] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 317–320.

[26] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[27] M. Radfar, R. Dansereau, and W.-Y. Chan, "Monaural speech separation based on gain adapted minimum mean square error estimation," *J. Signal Process. Syst.*, vol. 61, pp. 21–37, 2010.

[28] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc. Math Challenges of the 21st Century*, 2000, pp. 1–33.

[29] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. Eur. Signal Process. Conf.*, Aug. 2008.

[30] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proc. Int. Conf. Inf., Commun. Signal Process.*, 2005, pp. 1466–1470.

[31] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[32] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, pp. 249–252.

[33] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[34] P. Korten, J. Jensen, and R. Heusdens, "High-resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 966–981, Mar. 2007.

[35] P. Mowlaee, M. G. Christensen, Z. H. Tan, and S. H. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," in *Rec. Asilomar Conf. Signals, Syst., Comput.*, 2010, pp. 538–541.

[36] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[37] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion," *J. Acoust. Soc. Amer.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[38] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech dft coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Lett.*, vol. 15, pp. 213–216, 2008.

[39] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[40] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC, 2007.

[41] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 610–619, Feb. 2012.

[42] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel separation performance in transform domain," *J. Zhejiang Univ.-SCIENCE C, Comput. Electron.*, vol. 11, no. 3, pp. 160–174, Jan. 2010.

[43] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[44] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Speech Commun.*, vol. 2, pp. 749–752, Aug. 2001.

[45] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 561–564.

[46] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4214–4218.

[47] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. San Diego, CA: Academic, 1997.

[48] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[49] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans.Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[50] J. Ma and P. C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Commun.*, vol. 53, no. 3, pp. 340–354, 2011.
[51] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1–2046–2057, Sep. 2011.
[52] ITU-R BS.1534-1, Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems 2001.
[53] K. Krishnamoorthy, *Handbook of Statistical Distributions With Applications*. Boca Raton, FL: CRC, 2006, Univ. of Louisiana.
[54] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, no. 5, pp. 402–417, 2007.
[55] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.
[56] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Kinnunen, T. Fränti, and S. H. Jensen, "Sinusoidal approach for the single-channel speech separation and recognition challenge," in *Proc. Interspeech*, 2011, pp. 677–680.
[57] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010.

Pejman Mowlaee (S'07–M'11) received the B.Sc. and M.Sc. degrees both with distinction from Guilan University, and Iran University of Science and Technology, in 2005 and 2007, respectively. He received his Ph.D. degree at Aalborg University, Denmark in 2010. He is now a postdoctoral fellow at Institute of Communication Acoustics, Ruhr Universität Bochum. His research interests include digital signal processing theory and methods with application to speech, in signal separation & enhancement.

Rahim Saeidi (S'09–M'12) received M.Sc. degree in electrical engineering from Iran University of Science and Technology, Iran, in 2005, and Ph.D. degree in computer science from University of Eastern Finland, in 2011. He is a Marie Curie postdoctoral fellow working for EU funded BBfor2 ITN at Centre for Language and Speech Technology, Radboud University Nijmegen, Netherlands. His research interests include robust speaker recognition, speech enhancement, machine learning and pattern recognition.

Mads Græsbøll Christensen (S'00–M'05–SM'11) received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 2002 and 2005, respectively. He is currently an Associate Professor in the Department of Architecture, Design and Media Technology. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 1999. He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Denmark. His research interests include speech recognition, noise robust speech processing, multimedia signal and information processing, multimodal human computer interaction, and machine learning.

Tomi Kinnunen (M'12) received the M.Sc. and Ph.D. degrees in computer science from the University of Joensuu, Finland, in 1999 and 2005, respectively. From August 2005 to June 2007, he worked as an associate scientist at the Institute for Infocomm Research (I2R), Singapore. Since 2010 he has been employed as a postdoctoral research fellow at the University of Eastern Finland (UEF). His research areas cover text-independent speaker recognition and speech signal processing.

Pasi Fränti (M'00–SM'08) received the M.Sc. and Ph.D. degrees in computer science from the University of Turku, Finland, in 1991 and 1994, respectively. From 1996 to 1999, he was a Postdoctoral Researcher with the University of Joensuu, Joensuu, Finland (funded by the Academy of Finland), where he has been a Professor since 2000. His primary research interests are in image compression, clustering, speech technology and mobile location-based applications.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. He is Full Professor and is currently heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.